Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Computer Science 19 (2013) 1028 – 1033
Procedia Computer Science

The 3rd International Symposium on Frontiers in Ambient and Mobile Systems 2013 (FAMS 2013)

A Workload Analysis of Live Event Broadcast Service in Cloud

Yeali S. Sun[a], Y-F Chen[a], Meng Chang Chen[b]

[a]Department of Information Management, National Taiwan University
[b]Institute of Information Science, Academia Sinica

Abstract

The class of popular Internet live events, such as sports game broadcast services, can greatly benefit from cloud technology that employs resource pooling and rapid elasticity in resource allocation and management. The external demands tend to be unpredictable and exhibit a high degree of burstability. To sustain a good viewing experience for the users, an important QoS support for this class of applications is the peak load management in the presence of unpredictable demand. This relies on a close grasp of the demand behavior characteristics and an accurate prediction model for them. In this paper, we analyze live sports event broadcast service workloads from a commercial Internet service provider. Our results show that popularity follows a 2-mode Zipf distribution. We also observe some significant characteristics of the demand behavior. First, the demand behavior may differ significantly between games. Popular events tend to exhibit highly variable behaviors in time, volume and the change rate during the course. Finally, the demand variation highly correlates with certain event-specific time points. We conclude the paper with a preliminary study that applies three simple statistical models in workload prediction at runtime as an input to dynamic resource resizing in a cloud. The results indicate that more effective ways are needed to better capture the dynamics and unpredictability of the workload to improve prediction accuracy.

Selection and peer-review under responsibility of Elhadi M. Shakshuki

Keywords: live event broadcast service; workload analysis; dynamic resource allocation.

1. Introduction

There is a class of popular Internet applications which we consider can greatly benefit from the cloud technology, particularly the resource pooling and rapid elasticity capabilities in dynamic resource allocation and management. They are live event such as sports game broadcast service over the Internet.

This work was supported in part by NSC under grantsNSC 100-2219-E-002 -024 and 101-2221-E-001-021-MY3.

1877-0509
Selection and peer-review under responsibility of Elhadi M. Shakshuki
doi:10.1016/j.procs.2013.06.143

Different from on-demand shared media services, although this class of services typically have pre-scheduled event date and time, the external demands tend to be unpredictable and exhibit high degree of burstability in the way that they may encounter a sudden surge of viewer demand along as the event progresses. To sustain a good viewing experience for the service users, a cloud service provider should prevent such services from violating performance targets regardless of any change in external demand by utilizing the adaptive resource resizing capability of the cloud. Peak load management in the presence of unpredictable demand relies on a close grasp of the demand behavior characteristics and an accurate prediction model of them.

There are extensive studies of traditional Web workloads (e.g., [2, 3]) and on-demand streaming media workloads (e.g., [4, 5]) on the Internet. In [5], the authors pointed out that in high-bandwidth network environments, a performance bottleneck of video connections is likely to be pushed back closer to the server. Hence, the design of the server systems plays an important role in delivering a good viewing experience of the steaming media services over the Internet. In these studies, pre-recorded, stored streams (e.g., news clips) are targeted. In contrast, there are only a few works on live streaming, such as [6], which studied the workload characterization of a live streaming service in Brazil. The focus of this study was on the analysis of the behavior of different users interacting in this type of service, and not the traffic characteristics. One interesting point they found in the study is that live media access workloads are likely to be highly dependent on the nature of the live content with strong temporal correlations, which are consistent with our findings. Over the past decade, the service and the content objects over the Internet have changed considerably. Our present work investigates an Internet broadcast service of live NBA games. We hope that this work can provide new information and insights for server system design so that this class of Internet applications can take full advantage of the dynamic resource allocation capacity of the cloud in performance management.

In [7], a data set of the live streaming workload from Akamai was analyzed. However, in this data set, only 1% of the requests were for recorded video streams with the remaining consisting of audio content. Moreover, many of the data were short and recurring. In their analysis, the authors observed that a large number of streams, including both audio and video, exhibit flash crowd behavior with the peak of 40 arrivals per minute. In our study, we observed a similar phenomenon but with a much higher peak rate of up to 3000 arrivals per minute. In [8], they presented traffic characterization of YouTube on an edge network; but this again focused on on-demand stored media streaming service.

The rest of this paper is organized as follows. In Section 2, we present the workload analysis. Section 3 shows the prediction accuracy of three simple statistical models, none of which are able to capture the monitoring and control model. Finally, section 4 gives the conclusion.

2. LIVE STREAMING WORKLOAD

2.1. Sources of Workload

The trace set logs the requests of the 60 NBA games on hiChannel [9] between February 15, 2012 and April 22, 2012. The free video/audio platform hiChannel operates in Taiwan and provides live broadcast services of various pre-scheduled sports events. This channel is owned by Hinet, which is not only the largest ISP in Taiwan but also a subsidiary of Chunghwa Telecom, the largest telephone company in Taiwan. The service employs the Microsoft Media Server solution; and four resolution types are currently supported (low, medium, high and HD of 600, 1200, 2000 and 5000 bps, respectively). The service is delivered in two ways: over the Internet as well as the ADSL residential network of Chunghwa Telecom. Whereas Internet users only have access to low resolution, users on the HiNet ADSL residential network can watch live media streaming at higher resolution for better viewing quality (depending on the access speed subscription).

2.2. Characteristic Observations

From the traces, certain significant characteristics were observed which we consider important for accurate demand profiling and prediction in dynamic resource resizing and allocation in the support of cloud application performance management. Due to the space limitations of this study, in Fig. 1 we show a representative request rate behavior of the games, which are the 12 game events in March, 2012. One can see that these games exhibit rather drastic demand patterns. Moreover, several salient points of big rises and falls are usually situated around certain *event related* time points, e.g., a consistent surge in the request rates occur *before* the games begin. There are also unexpected bursts that may occur during the course of the game due to some unexpected game activities that develop. For example, in the event of March 7, 2012, the game of New York Nicks vs. Dallas Mavericks, a sudden rise in the request rate near the beginning of the game was completely unexpected by the service provider. In this game, Knicks' player Jeremy Lin, of Taiwanese descent, was the center of immense public attention in Taiwan. When more viewers joined the service and shared their comments through social media like Facebook more viewers were attracted and joined the service. This is an example of the uncertain and unpredictable nature of this class of applications. The unexpected crowd of demands may also be caused by such uncontrollable factors as popularity and the rich-get-richer effects [10]. These conditions can pose a great challenge to the service provider to provide the sufficient resources to sustain the QoS of the service. Moreover, in many games, game events may undergo rapid changes and exhibit high variation from start to finish. This suggests that the next state of the demand may not be predictable merely based on past observations. This potentially high dynamism and variability of the demand of the target services render predictions problematic.

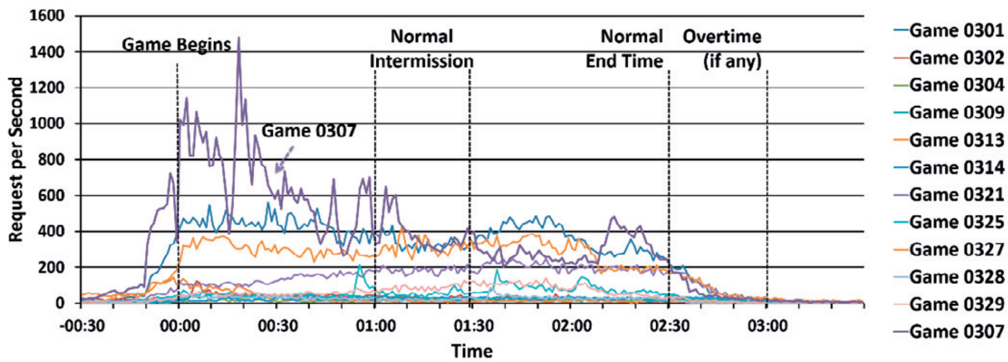


Fig. 1. Request arrivals of 2012 NBA live event broadcast service of hiChannel.

2.3. Characterization

In this section, we study the characteristics of the workload of this class of service.

A. Popularity vs. Requests - Zipf and Zipf-like Distribution

We define popularity as the total number of request arrivals of an event. In the literature, studies have shown that the relative frequency with web pages follows Zipf's law, such as in [3]. We were interested to determine whether the popularity distribution of the requests to online live sports event broadcast services similarly follows a Zipf-like distribution. Fig. 2 shows the total number of requests versus the popularity or rank of the 60 games; the observation period of each game event was 240 minutes starting thirty minutes before the game began. If the curve linearly fits the $\log(y) = -a*\log(x)+b$ where a is close to 1, then it follows Zipf's law. For a Zipf-like distribution, the value of a is typically less than one. The results show a 2-mode Zipf distribution which is consistent with studies of on-demand streaming objects [4, 11], multimedia file-sharing workloads [12], and the live content streaming workload of the mostly audio type [7].

We examine the cumulative distribution function (CDF) of the number of requests per game event in the traces as shown in Fig. 3. In Fig. 3(a), we observe a "27/73" rule which is less steep than the 20/80 rule. In Fig. 3(b), we fit all 60 and the top 42 games to a Pareto distribution and the values of shape parameter are 0.22414 and 0.72081, respectively. For the Pareto distribution, the value of the shape parameter is greater than 0; when it is less than one, it has an infinite mean. This has an important implication for resource allocation for server system capacity management.

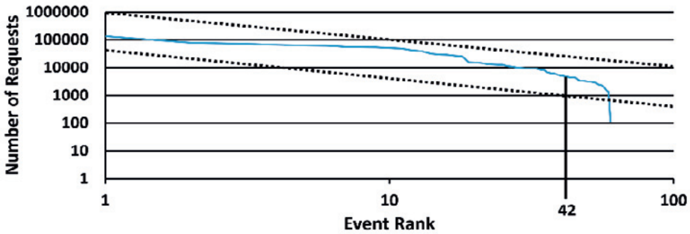
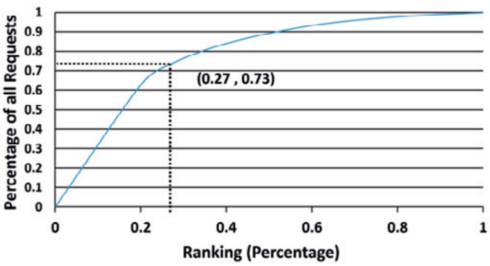
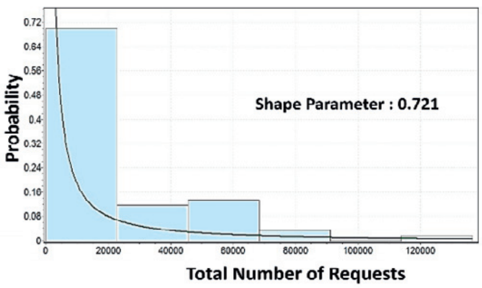


Fig. 2. Two-mode Zipf distribution of the 2012 NBA live event broadcast service of hiChannel.



(a) Fit for the "27/73" rule



(b) Fit for the Pareto distribution

Fig. 3. Some characteristic properties of the workloads.

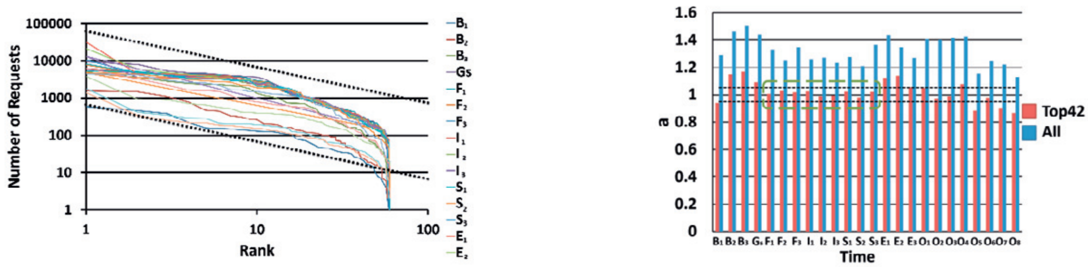
B. Game Periods vs. Requests – Zipf Distribution

Next, we analyze some temporal statistical and distributional properties of the workload. We divide the observation period of 240 minutes into 24 game periods. As shown in Fig. 4(a), they all follow 2-mode Zipf distribution. The values of a's are depicted in Fig. 4(b) for the sets of all 60 events (set ALL) and the top 42 events. We note that the periods following the Zipf distribution are during the course of the game for setTop42.

C. Burstiness vs. Popularity

We distinguish the 24 game periods into three zones: Z_1 of periods $F_1 \sim S_3$, Z_2 of periods $B_2 \sim G_s$ and

Z_3 of periods E_1 ~ E_3. Table 1 presents the comparison of zone 1, 2 and 3 in terms of burstiness and coefficient of variation (CV) for the three sets of traces of set ALL, set Top42 and set Top17 events(following 27/73 rule) . The set Top17 includes the five NBA finals and the 12 games with Jeremy Lin that attracted many viewers in Taiwan. We define burstiness as the ratio of average peak arrivals and average arrivals of the games during the periods in the zone. For the set ALL, the periods in Z_1 have the value of CV close to one while the periods in Z_2 and Z_3 have CV > 1, especially for Z_3. It is interesting that the more popular the event is the greater the dynamics exhibited. Moreover, the degree of fluctuation varies greatly in different zones.



(a) Two-mode Zipf distribution of game periods (b) values of *a* of game periods

Fig. 4. Some characteristic properties of the workloads of individual game periods.

Table 1. Comparisons of burstiness and CV of request arrivals in different time zones of the game events of different popularities.

	Z_1			Z_2			Z_3		
	ALL	Top42	Top17	ALL	Top42	Top17	ALL	Top42	Top17
Burstiness	5.85	4.27	2.20	7.11	5.15	5.93	15.80	11.47	5.66
CV	1.23	0.95	0.43	1.41	1.12	0.62	2.13	1.78	1.23

3. Prediction: Simple Statistical Models

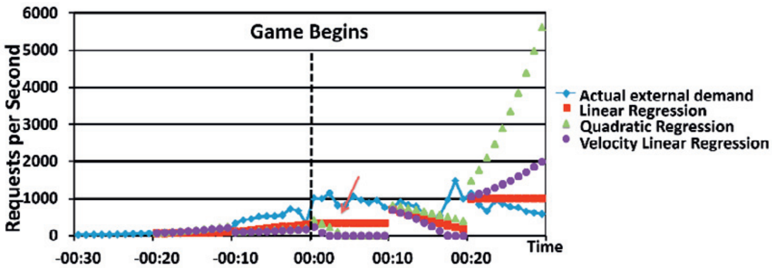


Fig 5. The comparison of actual and forecasted request rates using different statistical models

To support dynamic resource allocation in the performance management for cloud applications, we assume that a real-time demand measurement and dynamic resource allocation model is used. Periodic checks and controls (C&C) are performed for demand assessment and resource reallocation. We apply three simple statistical models for periodic demand prediction of the target class of applications: the linear trend model, the quadratic equation model, and a linear regression model to approximate the trend component of the rate changes. Fig. 6 shows the predication results of the three models. One can see that when the actual demand behavior of the next forecast period greatly differs from the previous history, these methods fail to give good prediction results. Other models that can better capture the workload changes and dynamics are needed.

4. Conclusion

In this paper, we present a workload analysis of the traces of an Internet live event broadcast service. We demonstrated that popularity follows a 2-mode Zipf distribution. We also observed some significant characteristics of the demand behavior. First, popular events tend to exhibit highly variable behaviors in time, volume and the change rate during the course of the event. Specifically, the periods in the event with high demand dynamics and variations are tightly correlated to certain event-specific time points, with different zones of game periods exhibiting quite different characteristics of request arrivals. This work of understanding the characteristic workload behavior of live event broadcast service is of fundamental importance in the design, operations and evaluation of Internet live content delivery server systems. We also showed that the intricacies of live event demand behaviour are unlikely to be well-captured by simple statistical models. A more effective demand prediction model is necessary to support target performance management and we are currently working on it.

References

[1] M. Ferrar, M. Gray, and K. Craig-Wood. Data Centre Migration, G-cloud and Applications Store Programme Phase 2: Technical Architecture Workstrand Report. Cabinet Office, 2010.

[2] M. F. Arlitt and C. L. Williamson. Internet Web Servers: Workload Characterization and Performance Implications. *ACM/IEEE Transactions on Networking*; 5(5):631–645, October, 1997.

[3] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In Proceedings of the IEEE INFOCOMM '99, March 1999.

[4] M. Chesire, A. Wolman, G. Voelker, and H. Levy. Measurement and Analysis of a Streaming-Media Workload. In Proceedings of Usenix Symposium on Internet Technologies and Systems (USITS), March 2001.

[5] Y. Wang, M. Claypool, and Z. Zuo. An Empirical Study of RealVideo Performance Across the Internet. In Proceedings of Internet Measurement Workshop (IMW), November 2001.

[6] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A Hierarchical Characterization of a Live Streaming Media Workload. In Proceedings of Internet Measurement Workshop (IMW), November 2002.

[7] K. Sripanidkulchai, B. Maggs, and H. Zhang. An Analysis of Live Streaming Workloads on the Internet. In Proceedings of Internet Measurement Workshop (IMW), October 2004.

[8] P. Gill, M. Arlitt, Zongpeng Li, and Anirban Mahanti. YouTube Traffic Characterization: A View From the Edge. In Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference (IMC'07), October 2007.

[9] http://hichannel.hinet.net/.

[10] David Easley and Jon Kleinberg Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010.

[11] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon. Analysis of Educational Media Server Workloads. In Proceedings of NOSSDAV, June 2001.

[12] K. Gummadi, R. Dunn, S. Saroiu, S. Gribble, H. Levy, and J. Zahorjan. Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload. In Proceedings of ACM SOSP, October 2003.